Environment and Pollution; Vol. 2, No. 1; 2013
ISSN 1927-0909 E-ISSN 1927-0917
Published by Canadian Center of Science and Education

A Semi-parametric Regression Model to Estimate Variability of NO_2

Mieczysław Szyszkowicz[1], Mamun Mahmud[1] & Neil Tremblay[1]

[1] Health Canada, Population Studies Division, Ottawa, Canada

Correspondence: Mieczysław Szyszkowicz, Health Canada, Population Studies Division, 269 Laurier Avenue, Ottawa K1A 0K9, Canada. Tel: 1-613-946-3542. E-mail: mietek.szyszkowicz@hc-sc.gc.ca

Received: July 9, 2012 Accepted: August 9, 2012 Online Published: December 4, 2012

doi:10.5539/ep.v2n1p46 URL: http://dx.doi.org/10.5539/ep.v2n1p46

Abstract

The purpose of this analysis was to derive a land-use regression (LUR) model using a semi-parametric method (based on penalized splines) to estimate the geographical characteristics that influence ambient concentrations of nitrogen dioxide (NO_2) in Montreal, Quebec, Canada. Such estimations are often used to assess exposure to traffic-related pollution in epidemiologic studies. In May 2003, levels of NO_2 were measured for 14 consecutive days at 67 sites across the city, using Ogawa passive-diffusion samplers. Concentrations ranged from 4.9 to 21.2 ppb (median 11.8 ppb). This work is re-analyzing of these data. Linear and semi-parametric multivariate regression analyses were conducted to assess the dependency between logarithms of concentrations of NO_2 and land-use variables. In the published multiple linear regression analyses for this study, distance from the nearest highway, length of highways and major roads within 100 m, traffic count on the nearest highway, and population density showed significant associations with NO_2. The best-fitting linear model had a R^2=0.54. The most important variable in the model was traffic count on the nearest highway. The next most important variable was distance from the nearest highway, which has a negative association with NO_2 concentration. This work used a semi-parametric model with a nonparametric part incorporating the variables "area of open space within 100 m" and "length of minor roads within 500 m". These variables were non-significant in the linear regression model and showed nonlinear associations with the level of NO_2. The semi-parametric model improves the fit of the model for land-use regression when comparing observed and predicted results.

Keywords: ambient air pollution, LUR, nitrogen dioxide, regression, road, traffic

1. Introduction

This work is an extension of the published analysis conducted by other researchers using the same data (Gilbert, Goldberg, Beckerman, Brook, & Jerrett, 2005). In their methodology, the authors of that publication applied only a linear multivariate model. This is a standard approach for geographic information system (GIS) modelling. Here, in this article, a semi-parametric model has been proposed. More recently, land-use regression (LUR) methods using GIS modelling have been developed and widely used. These methods have been used mostly to estimate the impact of traffic-related pollution on human health. The respective literature is presented in the original publication (Gilbert et al., 2005).

The current study was conducted to assess the feasibility of a large-scale monitoring campaign, in conjunction with land-use regression models, to estimate recent and past levels of pollution related to traffic. The results are applied in a major cancer case-control study that was conducted in Montreal over recent years.

In general, the land-use regression approach has two important features: a) building a good fit (LUR model) to the measured data (an interpolation process), and b) generating a good prediction (an extrapolation process). The prediction involves applying the model using a new set of data points. Thus, the process is to build a LUR model very specific to the measured data and to later use it to predict values of the new dataset. Using available measured data, it is possible to construct a model that provides the best fit for a given criterion. Mainly, it holds true when a semi-parametric approach is used and nonlinearity has been incorporated into the constructed model. A nonparametric part of the model can easily fit nonlinear relations in the measured data. From another point of view, apparently the classical linear land-use regression model ignores all nonlinearity in the data; therefore, nonlinearity and scatter smoothing may present a major opportunity to construct the best fit to the measured data. Such a precisely fitted model may not necessarily be a good one as a universal predictor. In other words, good interpolation for given data is not necessarily best for the extrapolation process for a new set of data. Rather, some balance and proper interpretation of variables should be done before accepting the specific model. This suggests that a semi-parametric method, a mixture of linearity and nonlinearity, may be a good approach to use in GIS modelling.

2. Materials and Methods

The methodology for monitoring nitrogen dioxide (NO_2) in Montreal is described by the authors in the original publication (Gilbert et al., 2005). In their study the monitors were set up at sites for a period of two weeks. Their paper refers to the distribution of NO_2 concentrations and the land-use variables selected, and also provide more information on the study. Here, to avoid repetition, many details are not included.

Different linear multivariate models were fitted to the original data. To assess the validity and the robustness of the approach, one-tenth of the data observations were selected systematically (e.g., 1st, 11th, 21st observation) and excluded. Thus, two models were used: one was fitted to the whole dataset (N=67 points), and then the model was fitted to the reduced data (N=60 points). The model obtained for the reduced data was used to predict values for the 7 removed points. This approach was applied in the original work (Gilbert et al., 2005).

A semi-parametric model was fitted using the *SemiPar* package from the R statistical system (Wand et al., 2005). *SemiPar* is free software and a simple tool to construct a nonlinear regression. In a univariate case, a fully nonparametric regression model has the form

$$y_i = f(x_i) + \varepsilon_i ,$$

where (x_i, y_i), $1 \le i \le n$, are the scatter plot data, ε_i are zero mean random variables with variance σ_ε^2 and $f(x) = E(y \mid x)$ is a smooth function. In the *SemiPar* package, *f* is estimated using penalized spline smoothing. To fit a semi-parametric model, the function *spm* (from the package *SemiPar*) was used. For example, the *fitlogNO*$_2$ model (which is linear with respect to the *distH* (distance from nearest highway) variable and nonparametric with respect to the *areaS* (area of open space within 100 m) variable) is constructed by invoking the following command:

$$fit\,log\,NO2 < -spm(log\,NO2 \sim distH + f(areaS)) .$$

A univariate, non-parametric model was constructed separately with each single independent variable used in the land-use multivariate linear regression model, to assess its nonlinearity in relation to the levels of NO_2. Two variables were classified to be in a nonparametric part of the semi-parametric model. The criterion of nonlinearity was used to classify the variables (Wand et al., 2005). A full semi-parametric model was developed using the same variables as in the linear regression model. Both models, linear and semi-parametric, were compared by assessing their fit to the full data.

3. Results

Table 1 presents the results of the multivariate linear regression. The results are the same as those obtained by the original authors (Gilbert et al., 2005).

Table 1. The results from linear regression models with N=67 and N=60 observations. The beta coefficients (Beta) are shown for the standardized data (N=67)

Covariates	Beta (N=67)	B (N=67)	p-value	B (N=60)	p-value
Intercept		0.745	0.000	0.757	0.000
Distance from nearest highway	-0.306	-0.0254	0.003	-0.026	0.004
Traffic count on nearest highway	0.358	$1.61\text{x}10^{-6}$	0.002	$1.56\text{x}10^{-6}$	0.004
Length of highways within 100 m	0.228	0.132	0.017	0.122	0.029
Length of major roads within 100 m	0.242	0.138	0.018	0.104	0.095
Length of minor roads within 500 m	0.183	$6.38\text{x}10^{-3}$	0.106	$7.32\text{x}10^{-3}$	0.087
Area of open space within 100 m	-0.184	-0.027	0.093	-0.020	0.272
Population density within 2000 m	0.284	$1.25\text{x}10^{-5}$	0.039	$1.11\text{x}10^{-5}$	0.079

The table represents the best fitting linear regression model with a determination coefficient R^2=0.54. The results are shown in the following columns: calculated coefficients (B) and p-values (p), both for the full data set (N=67 observations) and for the reduced data set (N=60 observations). Here, as an additional result, the beta coefficients (Beta) are shown for the full data set. The beta coefficients are the regression coefficients that have been calculated for measured data with standardizing all variables to a mean of 0 and a standard deviation of 1. Thus, the advantage of beta coefficients (as compared to the B coefficients that are not standardized) is that the magnitude of these beta coefficients allows for a clearer comparison of the relative contribution of each independent variable prediction to the dependent variable. The traffic count for the nearest highway (Beta=0.358) and the distance from the nearest highway (Beta=-0.306) are the two most important and statistically significant variables in the model (presented in Table 1). The results show that two variables (length of minor roads within 500 m, area of open space within 100 m) in the linear regression model are not statistically significant. In the fitted model for reduced data (only 60 observations), two additional variables are not statistically significant, demonstrating some sort of instability. This issue will be discussed later.

Figure 1 shows the results for a few variables used in a univariate nonparametric regression. The *SemPar* package provides information on the degrees of freedom for each fitted nonparametric element. This value can be used to classify the degree of nonlinearity (Wand et al., 2005).

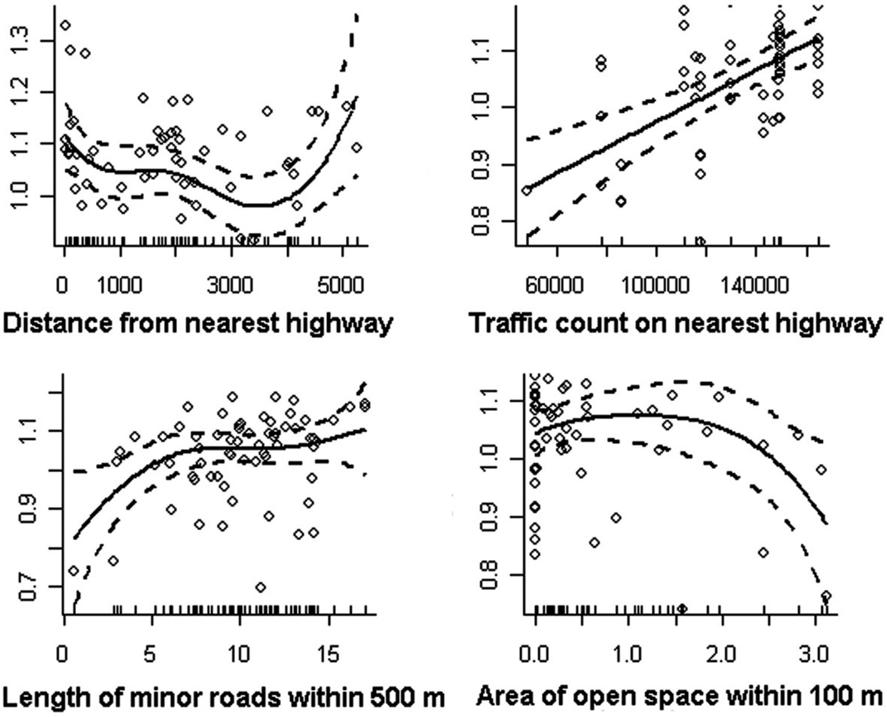

Figure 1. Semi-parametric models for $\log_{10}(NO_2)$ using single variable (a few chosen to illustrate)

Thus, potential candidates to be included in the nonparametric part of the semi-parametric model are: distance from nearest highway, length of minor roads within 500 m, and area of open space within 100 m. These variables in the univariate model have the following degrees of freedom (df): 4.2, 3.6 and 3.0, respectively. The value of df=1 indicates that the variable has a linear dependency; values greater than 1 suggest nonlinearity (Wand et al., 2005).

In the construction of a semi-parametric model, only two variables were included in its nonparametric part: length of minor roads within 500 m, and area of open space within 100 m. The results are shown in Table 2.

Table 2. The results from the semi parametric model with N=67 and N=60 observations

Covariates	B (N=67)	p-value	B (N=60)	p-value
Intercept	0.888	0.000	0.757	0.000
Distance from nearest highway	-0.026	0.002	-0.026	0.004
Traffic count on nearest highway	$1.66\text{x}10^{-6}$	0.001	$1.56\text{x}10^{-6}$	0.004
Length of highways within 100 m	0.107	0.044	0.121	0.029
Length of major roads within 100 m	0.124	0.028	0.104	0.095
Length of minor roads within 500 m	df=3.13		df=1.00	
Area of open space within 100 m	df=1.68		df=1.01	
Population density within 2000 m	$1.15\text{x}10^{-5}$	0.043	$1.11\text{x}10^{-5}$	0.079

The variable (with high df=4.2) corresponding to distance from nearest highway was not added to a nonparametric part. For this variable it was observed that, for values greater than 4 km, the level of NO_2 begins to increase. It suggests that there may have been another source of NO_2 during measurements that was not identified. For data restricted to distances of less than 4 km, the relation to NO_2 is linear.

In Table 2 the last two columns correspond to the subset of the data with 7 points removed. In this case, two nonlinear components in the fitted semi-parametric model start to become linear. Now the model is the same as that constructed by multivariate linear regression. Two last columns in Table 1 confirm this, and Figure 2 illustrates this phenomenon.

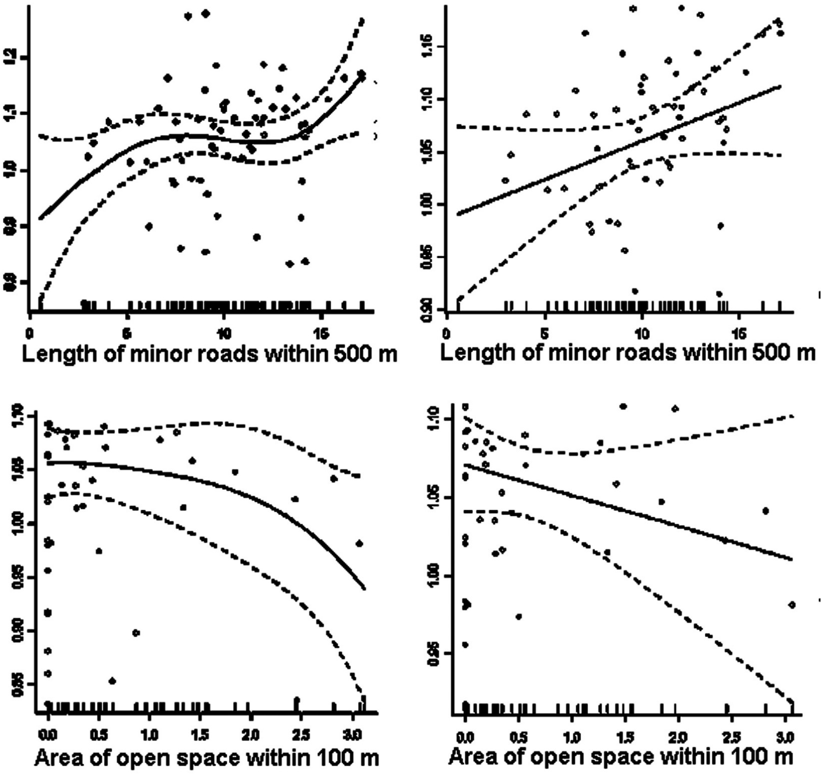

Figure 2. Two components of the semi-parametric model for $\log_{10}(NO_2)$ with N=67 points (left) and N=60 points (right). The results related to the values in Table 2

The removed 7 points are actually some specific points which strongly affect the regression model, and the fitted models are sensitive to these point.

The linear regression for the predicted NO_2 values was calculated for both models. These values are used as a dependent variable, while the actual measured values of nitrogen dioxide are used as an independent variable. For the original multivariate linear model, the coefficient for the measured NO_2 is 0.50, R^2=0.43, correlation=0.66. For the semi-parametric model, the coefficient is 0.56, R^2=0.53, correlation=0.72. This demonstrates that the semi-parametric model provides a better fit to the data than does a multivariate linear model.

3. Conclusions

In this paper it is shown that a semi-parametric model is a good alternative to a linear model. The main lesson is that the fitted models are very sensitive to the data. The dataset with N=67 points and its subset composed of N=60 points show a dependency characteristic that can change from nonlinear to linear. This change also emphasizes another problem: the best fitted model to the data is not necessarily a good tool (model) for an extrapolation process. The proposed semi-parametric model fits the original data well, adopts for nonlinearity, and starts to become equivalent to linear regression if the data do not show nonlinear relations.

Acknowledgments

The authors thank Nicolas Gilbert for providing the data related to this study.

References

Gilbert, N. L., Goldberg, M. S., Beckerman, B., Brook, J. R., & Jerrett, M. (2005). Assessing spatial variability of ambient nitrogen dioxide in Montreal, Canada, with a land use regression model. *J. Air & Waste Manage. Assoc, 55*, 1059-1063. http://dx.doi.org/10.1080/10473289.2005.10464708

Wand, M. P., Coull, B. A., French, J. L., Ganguli, B., Kamman, E. E., Staudenmayer, J., & Zanobetti, A. (2005). *SemiPar v.1.0*. R. 2.5.1. The R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/